CITY OF
SAN JOSE
CAPITAL OF SILICON VALLEY

Open Data Community Architecture (ODCA)

Version 3.2 (9/2018)

Authors

Arti Tangri

City Data Architect
arti.tangri@soanjoseca.gov

Rob Lloyd

Chief Information Officer
rob.lloyd@sanjoseca.gov

Partner

DELL EMC

SILICON VALLEY community foundation®

Table of Contents

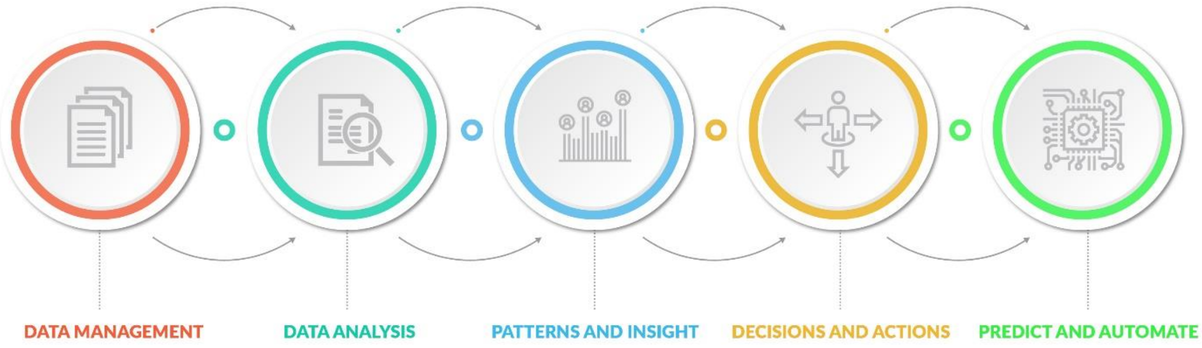
Executive Summary

The promise of "open data" grows from its ability to turn information into brilliant actions. This obvious but singular notion connects the Open Data Movement with the Smart Cities concepts that leading communities are embracing. It is a powerful commitment— superior understanding feeds exceptional insights and effectiveness at all levels of an organization. Those governments that build and enable the value of the data that courses through their people and systems are consistently positioned to harness information to better engage their community, help families thrive, and assist businesses to grow.

Yet, most would agree that the Open Data Movement has been stuck. The challenges of creating the human side of a data-driven organization are profoundly difficult. It demands knowledge and science that are inherently different from today's state of practice. Applications of Open Data that drive superior service or that yield significant savings are still exceptional events. Frameworks that fuel action, automation, and prediction are only now taking hold.

The purpose of this Open Data Community Architecture (ODCA) is to aid governments in driving toward those higher levels of decision maturity. The architecture builds around the notion that intelligent ecosystems reach optimum point when they are easily adopted, allow information to be shared broadly and openly, drive to smarter actions, and span communities, regions, states and nations. Thus, the ODCA marries an applicable reference architecture that communities can quickly use with the capacity to grow functions as they attain higher levels of maturity. This includes building data ecosystems that create incredible value through collaboration by communities of practice— governments, businesses, and academia.

The fact is that this work will be ongoing. The ODCA begins as the work of a few governments and leading companies. We expect it to evolve over time through the partnerships it encourages. And we hope you will be part of that work…



Concept

"Open data" refers to data that is not restricted. Legally, it is licensed for non-commercial use without restriction. Financially, it is free of cost. And technologically, it is available in machine and human readable forms that can be reused by various audience.

Like everything else, data comes in a variety of shapes, size, sources, and formats. Video versus text, streaming versus static, from private industry, government, non-profit organization or public.

Governments specifically have a very important role in open data. Public institutions collect vast amounts of data over time, serving as the repository for critical public records on which individuals and businesses must rely. According to a McKinsey report, open data can help create $3 trillion a year of value in the global economy through increased efficiency, promoting transparency, fostering competitiveness by making more information available, and creating opportunities to better match supply and demand. For example, open data has shown to provide significant value in transportation uses by providing real time traffic, parking, mass transit, and toll pricing along various routes to travelers. Making this data available to citizens helps them decide and act in more efficient ways, cutting the congestion, pollution, and transit time because fewer people opt for less efficient choices once they have information and easy alternatives.

Opening data within the City is the first step to making public Open Data more effective. In return it also helps make City operations run more efficiently and effectively. There are several data systems that have data that is used by different departments. If that data can be opened up to easily be consumed within the City, the net outcome of the combined datasets can be much more meaningful.

One key to the long-term success of the Open Data Community Architecture is how effectively it enable communities to share consumable Data at scale—what we refer to as Data Lakes feeding Data Oceans and/or Data Wells. A Data Well would be specific to a service to which City's Data Lake would feed. A Data Well would have Agreements, Access control and Data Standards that would set rules for sharing data across multiple organizations/agencies and would serve as a centralized data access point for data users. As an example, a cybersecurity Data Well could be shared with the cybersecurity experts in the industry and would comply with standards like TAXI and STIX. Similarly, Data Wells could be built around Disaster/Emergency Response, Transportation, Crime, City events and Homelessness.

A larger community that contributors can tap enables borderless traffic and crime analysis, as well as a new level of analytic solutions. Open Data initiatives will need to be able to connect ecosystems and tap a mix of market-driven data sources and tools. If the ODCA can make data more liquid and help get the local governments and businesses involved, data can contribute to achieving initiatives like the San Jose Smart City Vision.

Consumers can benefit from open data by gaining more insights in to what and from where they buy, where they go to school, and how they get around. Below is a short list of what the open data initiative at City of San José is aiming to achieve:

- Provide transparency by sharing data in a consumable form with the public.
- Empower individuals, media, civil society and businesses achieve better outcomes in public services.
- Help build a stronger, interconnected societies that meet the needs of the citizens and allow innovation and prosperity.
- Generate insights for better decision making internally and government-to-government.
- Build an ecosystem comprising of Academia, Businesses, Governments that uses a common data lake repository for data sharing.
- Maximize use and derive maximum value through high ecosystem adoption.

Scope

Open Data has a much broader scope than public data. Each city department has data that other departments can use and benefit from, but do not necessarily have access to. Heterogeneous data coming from siloed systems, lack of good visualization tools and lack of centralized data integration tools are some of the data challenges for the City that make it harder to get data together from multiple systems. The Open Data Community Architecture aims to provide a big data framework for data sharing within the City, with outside partners and with public at large. In addition to supporting data access within and outside the City, the Open Data Community Architecture also sets up a process and framework for data curation—from creation to archival/deletion. Eventually helping provide meaningful data to the end user, be it citizens, businesses or other government entities to help form a community that can benefit from sharing data, to ***transform data to decisions and actions.***

Goal

The goal of the Open Data Community Architecture (ODCA) is to create an integrated data platform that streamlines City's data flow by providing a mechanism for data aggregation, data sharing, and data analytics, thus leading up to predictive analytics. The data could be coming from legacy systems which could hold enormous amount of data dating back decades or even a century; or it could come from any kind of Internet of Things (IoT) device, ranging from sensors to cameras, which might not be that old but would still be enormous in size; the data could be structured or unstructured. As long as there is utility in the data to bring it together in a federated environment, to share it with intended parties, to combine it with other data, that data needs to be brought to a common data platform that this architecture outlines. The intent of the platform is to bring data from disparate systems in to one common location and provide a common data interface in the form of a self-service data access point for end users to use in analytics tools.

IoT Strategy

The data from IoT devices is becoming more and more important for analytics. Being able to bring that data to City's infrastructure is critical to eliminate any dependencies on third party vendors and to allow City to freely use this data. While not all of this data would be useful and the return on investment and the utility of this data would need to be evaluated based on specific requirements before bringing it in to City's infrastructure. Data streams from sensors and other IoT devices has two primary values. The first is for real time alerts and real time data driven dashboards that will allow the city to respond to incidents more effectively. The second is the accumulation of IoT data over time, which then enables predictive analytics.

Usage

The City sees this process in two primary phases beginning with leveraging data visualization tools to creating actionable intelligence from data to using predictive analytics tools to enable proactive resource allocation. As an example, Bill Schmarzo, in his article used accident data from City's open data portal to demonstrate how using analytics the data can be used for developing a self-learning AI solution. For modern Artificial Intelligence tools data size matters. Simple algorithms with lots of data will outperform sophisticated algorithms with less data.

The City intends to lead and support open source and open community approaches for data sharing, decision-making driven by a data-centric culture, and applied machine learning solutions to advance municipal services to improve citizen, resident, and visitor experience for data sharing through a scalable high-performance data lake platform.

DataOps

DataOps has gained popularity in last few years. As per Gartner's definition DataOps is "*a hub for collecting and distributing data, with a mandate to provide controlled access to systems of record for customer and marketing performance data, while protecting privacy, usage restrictions, and data integrity.*"

DataOps is an agile operations methodology that focuses on cultivating data management practices and processes that improve the speed and accuracy of analytics. It is about aligning the way you manage your data with the goals you have for that data. It is about democratizing the data and enabling every team to leverage data more effectively and efficiently.

Agile methodology is about making the best use of your time by utilizing something that has already been created. DataOps applies this concept to data. By setting up a common access point for the data we are avoiding having to provide access to the data for multiple use cases again and again.

As per the O'Reilly e-book Creating Data Driven Enterprise with DataOps, the key to success is becoming a data driven organization where your employees always use data to start, continue or conclude every single business decision, no matter how major or minor. There is a top-down push needed to initiate the process, but the real win is to have the bottom-up demand for self-service data access and for that to happen the self-service tools and processes need to be in place.

Development of Architecture and Resources

The Open Data Community Architecture (ODCA) is as much about design and architecture as it is about developing the supporting community of practice. To be successful, the citywide data architecture must develop use cases and resources that make the approach increasingly valuable.

Initially, ODCA work focuses on a highly adoptable reference build that all communities can use. From that starting point, contributors will work together to vet and extend the designs, creating standards for feeding the architecture's outcomes. As practices and technical standards shape, the critical step to solve for inter-organizational data sharing and usage will be tapping partners to create and support the underlying Data Lakes and broader Data Ocean that all will use. A few potential partners for the larger Data Ocean shared across multiple organizations are Microsoft Azure and Kaggle.

1. **Internal City Draft Architecture**
   - Detail City of San José's challenges and uses
   - Document a possible architecture to direct City investments and technical direction
2. **Contributing Partner Review and Validation of Architecture**
   - Validate and test technologies for a highly-adoptable, primarily open source architecture
   - Document technical components and options
   - Document integration and related needs to encourage vendor community to address
3. **Publication for Broad Peer and Vendor Review and Input**
   - Spur adoption to generate interest and additional testing
   - Develop additional uses cases to extend and test—e.g., open data model for disaster management, improved geospatial integration, et al
4. **Dedicate Architecture for Open Use**
   - Publish a finalized ODCA for all communities
   - Work with vendor partners and non-profits to develop resources for supporting standards
   - Work with vendor partners and non-profits to fund Data Lakes and Data Oceans resources

Data Governance

Master Data Management and Data Integration

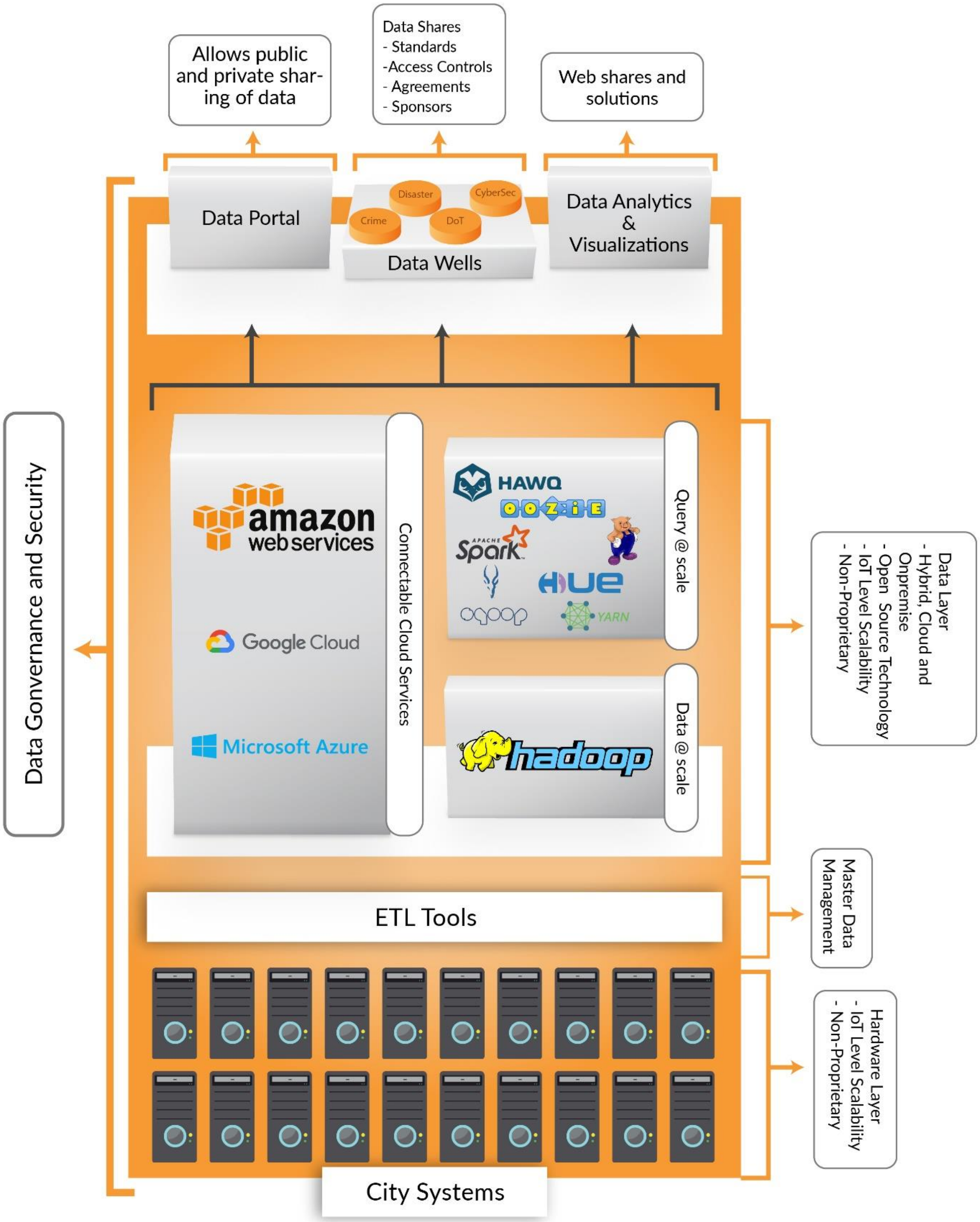
Master Data Management is about creating a single master reference source for all critical business data. Data integration is about efficiently managing data and making it available to those who need it. Collecting data from multiple sources while improving the quality of the data to give a single complete view of a business entity that is not possible when the data is dispersed. A few benefits of Master Data Management are as follows:

- *Providing a centralized location and making data more available* – As an example, a city maintains Address data. Address data is core to planning department as well as public works and housing. Other departments like Transportation and Parks & Rec also depend on it heavily. Having a common storage for Address data can be very helpful for all as it provides a single source of truth for Address information.

- *Reducing data complexity by having a single interface for data coming from different systems* – Adding on to the Address data scenario, each department uses and maintains different information from the Address data. To exchange information on Address with multiple departments would mean interacting with different systems and setting up different processes around information sharing. Having the same data in one place eases up the load of data collaboration.

- *Making data more valuable by combining different datasets together* - When the Address data from each department is combined together, be it zoning information from planning or infrastructure information from public works it makes the Address data much more valuable and complete.

- *Multiple city data set aggregation* - When the same type of data can be analyzed from multiple sources, the potential for data insight is increased. For example, air quality data analysis from several cities in the same region could lead to greater insights. This type of analysis could be more easily performed if some data models were standardized for this purpose. Air quality data could share a similar format between cities.


Allows public and private sharing of data
Data Shares
- Standards
-Access Controls
- Agreements
- Sponsors
Web shares and solutions
Data Portal
Disaster
CyberSec
Crime
DoT
Data Wells
Data Analytics & Visualizations
Data Gonvernance and Security
amazon web services
Google Cloud
Microsoft Azure
Connectable Cloud Services
HAWQ
OOZIE
Spark
HUE
sqoop
YARN
Query @ scale
hadoop
Data @ scale
Data Layer
- Hybrid, Cloud and Onpremise
- Open Source Technology
- IoT Level Scalability
- Non-Proprietary
ETL Tools
Master Data Management
Hardware Layer
- IoT Level Scalability
- Non-Proprietary
City Systems

Security & Privacy

The security strategy needs to address the prevention of cybercrime, data access control, and prevent risks associated with combining data sets to reach PII type conclusions. As an example, being able to strip out the personal information from crime data before publishing it would be very critical to ensure privacy of the public. There is also a need for Digital Privacy Policy that must drive the data sharing online.

While we are working with open public data it is important to realize that this data is coming from secured systems and when we get this data out open to the public we need to make sure we do not make our systems vulnerable in any way. One of the possibilities of achieving the security for internal systems is to have a staging data warehouse that extracts the data from other systems and then pushes it out to the data lake that then exposes the data to be consumed by users and applications. The data access through API and Micro-services should be the primary method of data sharing for the environment and would need to be secured. There needs to be an enterprise level security around the data lake environment to ensure reliability and security. One of the ways to achieve this would be to have an active directory integration.

This architecture needs to ensure that any information published online must not reveal the internal operations of the City which might put City at a risk of cyberattack. Any published data must not contain any personally identifiable information, the architecture must be governed by the federal, state or local privacy and data protection laws, whichever is applicable. This architecture intends to establish an internal review and approval process for data before it is made accessible outside. There should also be an option to have inter-department coordination for data sharing to ensure data common to multiple departments gets approved by all the departments to attain a city-wide level of accuracy.

Specifications

This section describes the high level requirements that the ODCA architecture is intended to meet.

1. **Cloud vs On Premise** – The goal of a city is to use the best tools for an overall information sharing solution. Some of these tools will reside in city data center and some will likely be hosted in the cloud. The decision whether to put a specific tool in the data center or cloud will be based on whether the tool has that option and the determination of what is most cost effective over time for the city. The cost analysis is a simple evaluation of comparing upfront costs with the total cost of ownership (TCO) over the expected life span of the system and at scale.

   Current expectations are that the core components of the ODCA would be located on premise in the city data center. If the city chooses to work with a cloud hosted open data provider, those data sets would exist in that location and may also exist in the city data center data platform. This will be determined based on the capabilities of the open data components of the ODCA architecture.

2. **ODCA City Data Lake Platform** – The city hosted data platform will be based on industry standard and open source tools and will allow for any type of data to be added. It will not require transformation prior to addition to the data lake. It is anticipated that the data could be both structured and unstructured and the system will allow for meta-tagging of the data or content such that it can be exposed through queries, dashboards and analytics tools. The data platform will primarily be based on a Hadoop data architecture and associated toolsets. The architecture will provide for a real time data engine capable of ingesting data streams and exposing this data to dashboards and analytics. It will also have a big data component which will manage the accumulation of data over time. The platform is currently anticipated to run in the city data center but could be moved to a cloud provider if that becomes the most cost effective model. For data sets which require

A few terms…

API: Application program interface (API) is a set of routines, protocols, and tools for building software applications.

[x]KAN: *Open source open data catalog systems*

Data Lake: a storage repository that holds a vast amount of raw data in its native format until it is needed.

Data Ocean: a storage of data from multiple data lakes.

ETL: Extract, Transform, Load, three database functions combined into one tool for moving data from one location to another.

Micro service: Componentization of applications via services

REST: Set of rules used by applications to communicate over the internet

Web Service: a standardized way of integrating Web-based applications using the XML, SOAP, WSDL and UDDI open standards over the Internet transformation, the system will allow for data to be extracted, transformed as needed and then re-added to the data lake

3. **Ease of Data Upload** – The system must make it easy for data sets to be added to the ODCA data lake platform. The plan is for data set uploads to be done by city agencies as well as by businesses, universities, and residents in the community. The system should allow a user to upload a data set either through an interactive web portal or programmatically through calling a RESTful Web Service API. Through user upload, the system should capture meta information that identifies the data set including meta tags such as date of load, responsible person, contact information, retention policy etc. This will allow for managing the content of the data lake and allow for the formation of information governance policies for the content. The system should require a minimum set of meta tags and reject data sets that do not meet the minimum required metadata. The data could also be a real time data stream from a device such as a sensor. This data stream would continue to feed data to the data platform once linked.

4. **Data Cleansing or Transforming** – The ODCA architecture should allow for the addition of these tools as needed. The general concept is that data can be uploaded and then extracted, transformed as needed and then re-added to the system. System administrators can then determine whether to keep both versions of the data or just the transformed or validated data set.

5. **Quarantine** – Any user uploaded data sets will be placed in a holding area or quarantine such that they can be evaluated prior to becoming available. Specific users or groups at the City will have administrative access to the quarantine holding area and will be notified by the system to review the data sets and ensure that there are no risks associated with making the information available to the receiving audience. Risks include Personally Identifiable Information (PII), SSN, Credit card numbers, and malware. Another risk is that this data set could be connected with other data sets and the result of that combination would create a PII (or other) risk. The system will not provide any automated risk analysis but will allow tools to operate against uploaded data sets to perform these evaluations. Once the content is deemed ready, it can be promoted to its end state access by the administrator.

6. **Security** – The security piece is taken care of in 3 stages – Authentication, Authorization and Access Control. The data platform will allow for a data set to be made available to anyone who has access to the system or only allow access to specific users or groups, taking care of authorization and access control depending on the level of access provided. For example, a city department may want to share a data set with another department but not have it available to other city departments or the public or the city department might want to share this with a department as a read only access or allow the other department to make edits too. This security access control should be enabled when the content is added to the system. Goal is also to integrate with Active Directory for city user authentication.

7. **Availability** – Once promoted, the data set will appear in the data catalog and be available to authorized users.

8. **Data Visualizations** – Data sets should be viewable or download-able in their native formats. Data visualization tools in the platform will allow for the creation of dashboards, graphics, and plotted on a map using geolocation data. Data will be accessible from 3rd party visualization and analytics tools as needed and through standard data access methods.

9. **Data Movement CKAN ←→ Data Lake** - Data Integration from CKAN to the Data Lake and exposing Data Lake data through CKAN. By using CKAN as the primary UI for data entry, upload, management, and access, we can leverage the capabilities existing in CKAN which provides for much of the requirements as described in this document. By also having the data moved or copied from CKAN to the Data Lake, the data will be made available to leverage visualization tools such as Tableau (or others) or other tools used for deeper analytics and machine learning. It is also anticipated that CKAN will be used to provide a data catalog of data sets by capturing the metadata of each data set and links to the data set in the data lake. In the event that the City desires to publish the data to a cloud based open data provider, the CKAN portal would still remain as the data catalog and would contain links to the published data set and the source data set in the data lake.

10. **Data API and/or Data Micro-service** – The system will allow for the creation of Data APIs and/or Data Micro-services which will allow other systems or applications to access the data sets or stream as a RESTful Web Service.

11. **Dashboards, Machine Learning, Real time Analytics, and Data Mining** – The system will provide access for both real time analysis and dashboards as well as predictive analysis of data. The system will provide for data access methods for these tools to function. The data could be real time data streams from sensors or other inputs feeding a dashboard, or it could be big data sets used for predictive data analytics.

12. **Data Sharing and Data Hosting for Others** – Once proven as a capable solution, the city desires to make the data lake platform available to others for use. This will help position the city as a leader in data sharing architecture and may provide some cost relief for the compute resources needed to power the system.

13. **System Infrastructure, Uptime, and Recovery** – The system will be implemented on a compute and storage environment capable of sustaining a big data architecture. It will also be architected such that it can be deployed with the following capabilities:

    - High Availability
    - Backup and Recovery
    - Disaster Recovery

14. **System Performance at Scale** – The system is intended to support the anticipated user volume and data volumes. As the data volume grows the system needs to not suffer performance degradation associated with large data volumes and increased user load.

15. **Containerization of the platform** – The success of the architecture lies in how easily it can be adopted by multiple agencies especially by those with only basic technical resources. Containerizing the platform allows for fast and easy deployment and better image management. Docker and Kubernetes are the two most popular methods of containerization available today. While containers provide a lot of benefits, its security is still in question. Until containers can provide a robust security, organizations need to work around this missing piece and create its own security to counter it.

Roles and Skills

People are at the center of this architecture. The focus of the architecture is to have an environment that can be easily used by the existing staff with minimal training for better and faster adoptability. The following are the roles that this architecture would need.

1. **Data Lake Administrator** – This role (shared by an existing team) would be centrally located in the IT Department, responsible for maintaining the data lake environment, troubleshooting the issues with servers and making sure all the applications are up and running. Skills required – Linux, CentOS, Windows, a high level understanding of Hadoop applications.

2. **Department Data Publisher** – A data publisher from each department will be responsible for uploading datasets in to the data lake environment. The data access to the data lake is meant to be via a UI tool and/or the API (directly or using an ETL tool), hence, no special skills are required for this role.

3. **Data Approvers** – This role is responsible for ensuring the data quality and would primarily access data via a UI tool. No special skills required.

4. **Data Consumer** – This is the end user that needs the data and would typically access it via an API. Skills required – Ability to access the data via an API.

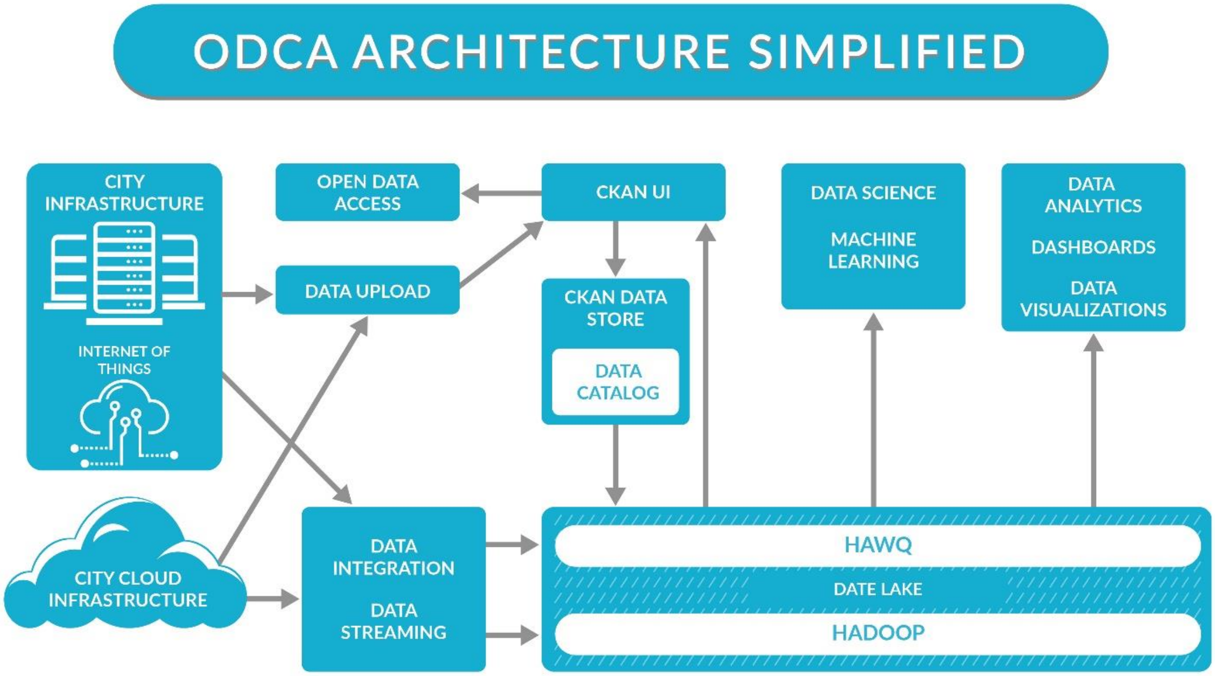
5. **Developer/Maintainer** – This would have to be an as needed external resource that would be doing any kind of enhancements and development work on the environment. Skills required – depends on the type of enhancements.

6. **Data Scientist** – In order to make use of predictive analytics tools, City needs to engage with Data Scientists directly, in combination with a local university, or through a service like Kaggle.

ODCA Reference Architecture

The diagram below shows the vision for how data will flow through the planned ODCA architecture. The components of the overall architecture will be implemented in phases with the first phase focused on the data lake components and open data portal. This includes the ability to add data to the data lake and expose and share data through the open data portal. The end state vision of the complete solution is attached as an appendix to this document.

Follow on phases will add the other components of the ODCA architecture as needed for Smart City projects.



The primary areas of functionality in the diagram are noted with numbers which will be described here:

1. **Data Ingest/Integration/ETL** – The solution anticipates that data will come from multiple sources and consist of many different types of data. The architecture will provide for a user interface where users can add data sets to the system through data upload. All data sets uploaded through the user interface will be quarantined until reviewed and validated by data administrators.

   The second type of data integration will be for automated data ingest for data. This could be real time data streams from devices such as IOT sensors. It could also be timed uploads from data providers. The key is that the architecture can support multiple data sets and data types, real time data streams, and big data.

One of the benefits of the data lake architecture is that data can be added in its native format and structure and does not require Extract Transform and Load (ETL) tools. In the event that data transformations or validations are needed, these tools can be run either as data is ingested or post ingest (ELT)

2. **Data Lake** – The primary data store for the architecture is a Hadoop based data lake. Hadoop is very capable of handling big data sets and stores but is not as well suited to real time data streams and associated real time dashboards and analytics. For these data streams, a real time data engine is required. The architecture leverages tools such as HIVE, Spark, or HAWQ as the real time data engine. Depending on the data type, size and frequency of update, the data will flow into the system either into the HDFS component or into the HIVE, Spark or HAWQ component.

3. **Data Management / Governance** – Components in this area of the architecture include the ability to create a catalog of data sets such that the system does not become a data swamp where data is added but not managed or governed. Data sets will be tagged with metadata and subject to data governance rules and policies including data curation. It is anticipated that CKAN will provide the data catalog component of the architecture.

4. **Exposing Data / Data Access / Data APIs** – The architecture anticipates that data will be exposed through standard data access methods. These standard access methods will allow for out-of-the-box data visualization tools. The data will also be exposed such that it can be accessed by more advanced data analysis tools such as machine learning and predictive analytics tools. Data can also be exposed as an API (RESTful Web Service) such that applications can be developed to leverage these data APIs. It is also possible to create the data API as using a micro services architecture for web scale applications.

5. **Data Usage** – Data can be exposed through the open data portal for data sharing either with the public or privately between city departments. Data can also be immediately used by out-of-the-box data visualization tools and dashboards. Real time dashboards can provide for actionable intelligence for city departments and management. Data as an API or service can be used for either city developed or third party developed applications. Also important is the value of big data for predictive analytics tools. The larger the sample size, the more accurate the predictive analytics can be. The architecture anticipates being able to leverage very large data sets generated from IOT and/or sensors over time.

Appendix A – Technologies for ODCA

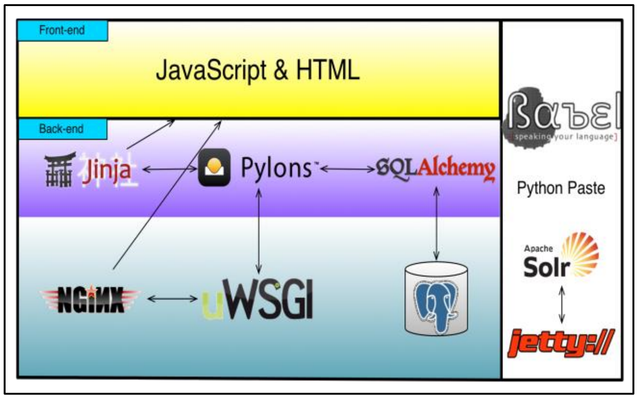
The following options were evaluated for architecture

Open source Data Portals

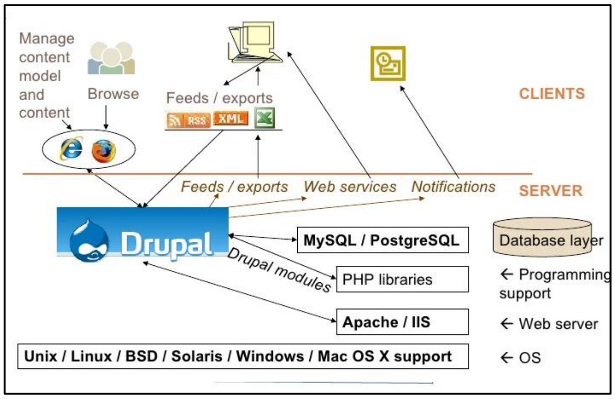
Function	CKAN	DKAN
Vendor/Sponsor	Open Knowledge Foundation	Nuams
Language used	Python and Python libraries	PHP
Tools used	Solr, Postgresql, Tomcat, Apache, JS	MySQL/SQL Lite, Apache, JS, Solr
Framework	SQL Alchemy, Pylons	Drupal
Content Management Framework Integration	No	Yes (Drupal)
Direct Data Access (File Store)	**No, it stores all the files in blob format and it will be accessed via API only**	**Yes, the files are kept in native format /var/www/html/dkan/sites/default/files**
Workflow	Basic workflow or publisher profile exists	Comprehensive workflow and approval flow exists
Extension	60+ Extension exists to support different features (e.g. Social Media, Data Export, etc.)	It will be supplied through Drupal Modules. No CKAN extension suited for DKAN
Maturity	High	Medium (Emerging one)
List of Open Data Portal across the world	200+	30+
Metadata/Data API	Yes	Yes
Metadata Integration	Yes	Yes
Support Community	Python developer community	Drupal developer community
Visualization	Yes	Yes
Dashboard	Yes	Yes

Function	CKAN	DKAN
Map Integration	Yes	Yes
Data Set (Format) Supported	XML, CSV, XLSX, JSON, KML, GEO JSON, GML, GFT, API	Same as CKAN
Security Integration	LDAP, Active Directory	Tightly integrates with Drupal support. Drupal supports LDAP and Active Directory)
Drag and Drop Layouts	No	Yes
Standards Complaint	DCAT	DCAT and US Project Open Data
Workflow Management	Basic	Advanced
Demo URL (Public)	http://demo.ckan.org/	http://demo.getdkan.com/
Level of Customization	Compared to DKAN, it supports less content modification	High level of customization through Drupal
Searching	Yes	Yes
Geographic Searching	Yes	Yes
Multilanguage support	Yes	Yes

CKAN



DKAN



Data Tools

The City evaluated Tableau and Microsoft Power BI for data visualization. Here's a quick summary of where each one stands.

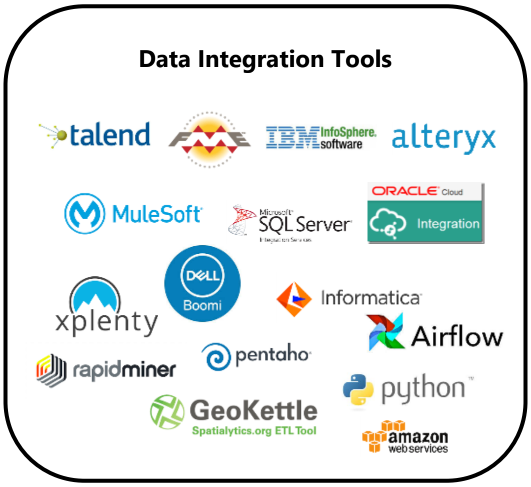
MS Power BI is a relatively newer product and is still evolving. It has an excel like look and feel which makes it more easily adoptable. It is heavily integrated with Microsoft's portfolio. Power BI is an economical tool and can be scaled depending on the use. Tableau is a more established product and has been around for much longer. It is easy to use, supports connectivity to distinct data warehouses and has the ability to integrate with infinite data points.

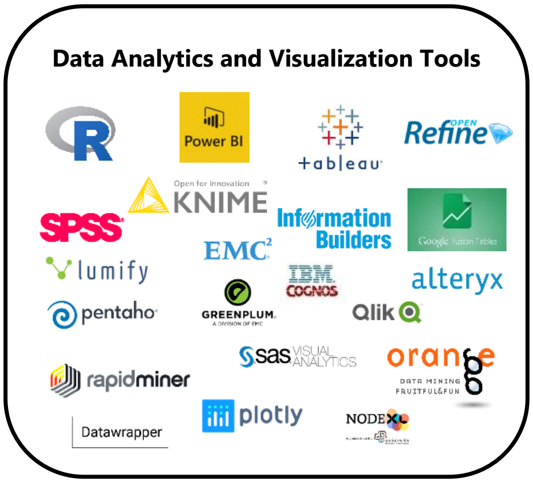
The City also evaluated Talend which is a free open source ETL tool and worked well with the architecture. Talend is a low-cost option, its portfolio includes data quality, MDM and API management tooling. Its tools are highly configurable which makes them flexible enough to adapt to changing business requirements. However, its user group is still small hence much of the support still relies on Talend.

There is a wide range of data tools that can potentially work with this architecture. Important factors to look for when evaluating the tools are

- Adoptability – The real success for the architecture is how well it gets adopted by the end users. The more user friendly and the wider the features are in the tools the more easily adoptable it is.
- Scalability – This architecture is designed to support big data and data from IoT devices, hence, it is extremely important that the data tools can scale up with the growth in data.

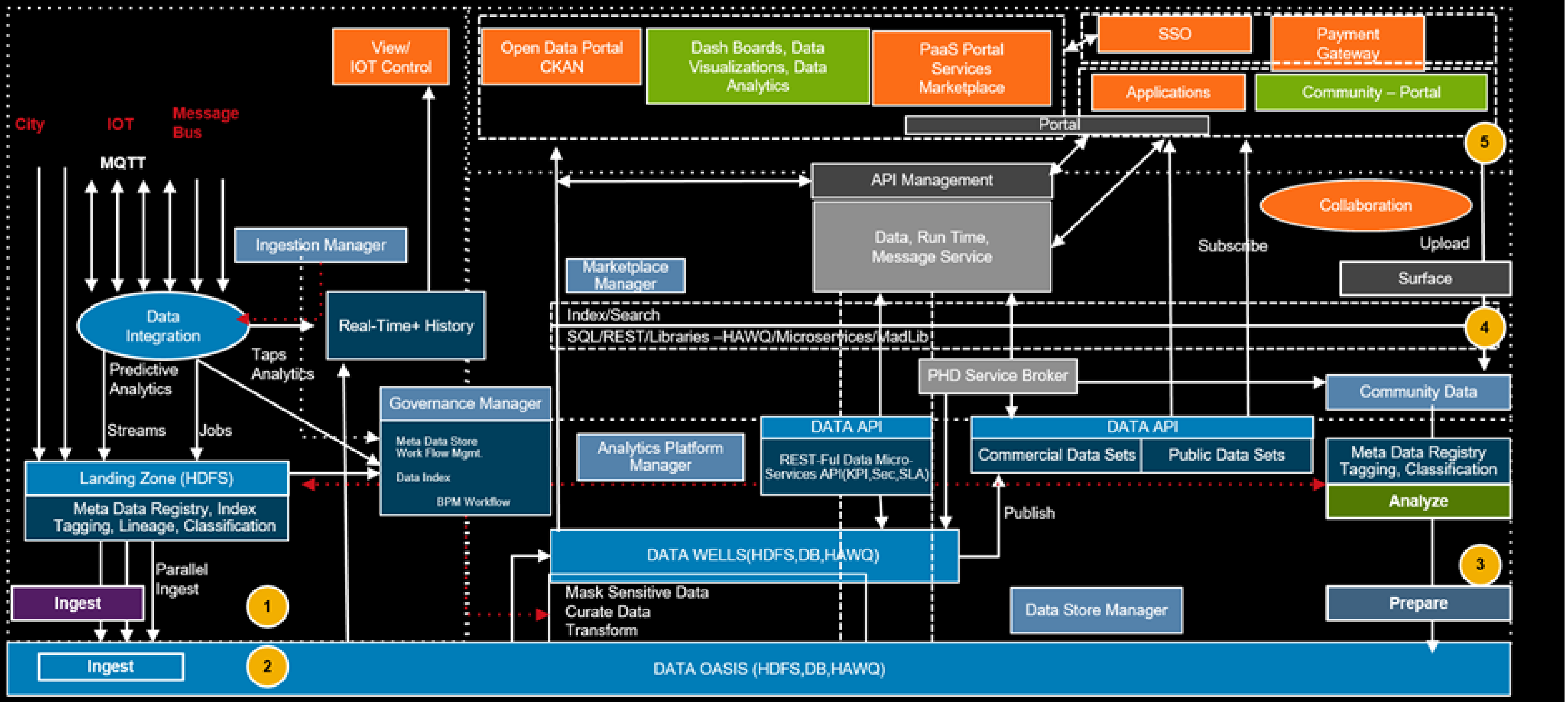
Some of the potential Data Integration tools and Data Analytics and Visualization tools are listed in the image below.





Appendix B – End State Architecture

San Jose ODCA Reference Architecture



City
IOT
Message Bus
MQTT
View/ IOT Control
Open Data Portal CKAN
Dash Boards, Data Visualizations, Data Analytics
PaaS Portal Services Marketplace
SSO
Payment Gateway
Applications
Community – Portal
Portal
5
API Management
Data, Run Time, Message Service
Collaboration
Subscribe
Upload
Surface
Ingestion Manager
Marketplace Manager
Data Integration
Real-Time+ History
Index/Search
SQL/REST/Libraries –HAWQ/Microservices/MadLib
4
Predictive Analytics
Taps Analytics
PHD Service Broker
Community Data
Governance Manager
Meta Data Store
Work Flow Mgmt.
Data Index
BPM Workflow
Streams
Jobs
Analytics Platform Manager
DATA API
REST-Ful Data Micro-Services API(KPI,Sec,SLA)
DATA API
Commercial Data Sets
Public Data Sets
Meta Data Registry Tagging, Classification
Analyze
Landing Zone (HDFS)
Meta Data Registry, Index Tagging, Lineage, Classification
Publish
Parallel Ingest
DATA WELLS(HDFS,DB,HAWQ)
3
Ingest
1
Mask Sensitive Data
Curate Data
Transform
Data Store Manager
Prepare
Ingest
2
DATA OASIS (HDFS,DB,HAWQ)

3

Validation

Third-party validation of this architecture is a crucial element to its continued viability. Following are the thought-leaders who have contributed to the review and the ODCA. Credit to these individuals for helping to ensure (1) applicability of the direction; (2) completeness of architecture, and (3) technical grounding that will support collaboration and machine-learning uses.

The ODCA is designed to support a community of practice for data use that supports communities. It is intended to be developed over time by experts in the field. Following is an accounting of contributors who have helped shape the document through its versions.

Log of Contributors:

Name	Representing	Title	Email	Date
Arti Tangri	City of San José	Data Architect	arti.tangri@sanjoseca.gov	4/30/2017
Rob Lloyd	City of San José	CIO	rob.lloyd@sanjoseca.gov	4/30/2017
Rob Silverberg	Dell EMC	CTO	rob.silverberg@dell.com	8/11/2017
OpenGov				March 2018
ACLU-Santa Clara				April 2018